

## Review of NLP in EMR: abbreviation, diagnosis, and ICD classification

Nurul Anis Balqis Iqbal Basheer<sup>1</sup>, Sharifalillah Nordin<sup>1</sup>, Sazzli Shahlan Kasim<sup>2</sup>, Azliza Mohd Ali<sup>1</sup>,  
Nurzeatul Hamimah Abdul Hamid<sup>1</sup>

<sup>1</sup>Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

<sup>2</sup>Cardiovascular Advancement and Research Excellence Institute (CARE Institute) Universiti Teknologi MARA, Selangor, Malaysia

### Article Info

#### Article history:

Received Oct 24, 2024

Revised Jan 2, 2025

Accepted Jun 9, 2025

#### Keywords:

Diagnosis

EMR

Expanding abbreviations

ICD classification

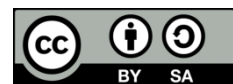
NLP

Review

### ABSTRACT

This review explores state-of-the-art natural language processing (NLP) methods applied to electronic medical records (EMRs) for key tasks such as expanding medical abbreviations, automated diagnosis generation, international classification of diseases (ICD) classification, and explaining model outcomes. With the growing digitization of healthcare data, the complexity of EMR analysis presents a significant challenge for accurate and interpretable results. This paper evaluates various methodologies, highlighting their strengths, limitations, and potential for improving clinical decision-making. Special attention is given to abbreviation expansion as a crucial step for disambiguating terms in the clinical text, followed by an exploration of auto-diagnosis models and ICD code assignment techniques. Finally, interpretability methods like integrated gradients and attention-based approaches are reviewed to understand model predictions and their applicability in healthcare. This review aims to provide a comprehensive guide for researchers and practitioners interested in leveraging NLP for clinical text analysis.

*This is an open access article under the [CC BY-SA](#) license.*



### Corresponding Author:

Sharifalillah Nordin

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA

Shah Alam, Selangor, Malaysia

Email: sharifalillah@uitm.edu.my

## 1. INTRODUCTION

Electronic medical records (EMRs) serve as comprehensive digital repositories for patient data, centralizing patient demographics, medical histories, test results, and treatments [1]-[4]. While EMRs have transformed healthcare by streamlining data management, they also introduce significant challenges due to the sheer volume and complexity of the data. Healthcare providers are often overwhelmed by the need to navigate and update these extensive records, leading to cognitive overload and potential burnout [5]. Furthermore, EMRs contain a large amount of unstructured data, which poses difficulties for tasks like automated diagnosis generation and international classification of diseases (ICD) classification critical for patient care, billing, and healthcare management [6]-[8].

The ICD is a globally recognized system for coding diseases and health conditions, playing a vital role in clinical decision-making, health management, and epidemiology [9]-[13]. However, assigning accurate ICD codes often involves manual input from medical coders, which introduces the risk of human error and inefficiency. Misinterpretation of clinical notes, particularly unstructured data, can lead to coding inaccuracies, compromising billing accuracy and healthcare quality [3], [14]-[17].

Relying solely on structured data in healthcare overlooks the complexity and richness of patient information, as it is limited to predefined categories like diagnosis codes and lab results. This can miss

critical context found in unstructured data, such as detailed symptom descriptions, treatment plans, and progress notes, which are vital for personalized care and accurate diagnoses. Additionally, the pervasive use of medical abbreviations in EMRs poses unique challenges, as their meanings often depend on context. Misinterpretation can lead to incorrect diagnoses, improper ICD classifications, and compromised patient outcomes, highlighting the need to integrate structured and unstructured data for comprehensive decision-making.

The best approach, therefore, is to combine both structured and unstructured data. Leveraging natural language processing (NLP) techniques allows healthcare providers to integrate structured, easily accessible information with the rich detail found in unstructured data. This combination enables a more comprehensive view of patient health, enhances diagnosis accuracy, improves treatment plans, and ultimately leads to better patient outcomes. Combining both data types also improves the efficacy of automated systems, such as ICD classification and diagnosis generation, making healthcare more intelligent and data-driven.

NLP offers powerful methods to address these issues. NLP techniques can process and transform unstructured clinical text into actionable insights, enabling more efficient data handling, supporting clinical decision-making, and improving patient care [18]-[20]. In this review, we focus on NLP solutions for four critical EMR-related tasks: expanding abbreviations, generating automated diagnoses, assigning ICD codes, and interpreting model predictions.

One major challenge in EMR interpretation is the pervasive use of medical abbreviations, which can vary greatly depending on context. Misinterpreting abbreviations can result in incorrect diagnoses or improper ICD classifications, ultimately affecting patient care and healthcare data integrity [21], [22]. Expanding these abbreviations accurately is critical to ensure the reliability of diagnoses and ICD assignments.

Traditional machine learning methods, such as support vector machines (SVM) and random forests, have been applied in this field, but their reliance on structured data makes them less effective for processing unstructured clinical notes, which require extensive feature engineering and domain expertise [19]-[22]. Newer NLP methods, however, are well-suited to handling unstructured data in EMRs and can improve efficiency by automating these tasks.

Although research has focused on each of these areas independently, limited work has integrated abbreviation expansion, automated diagnosis, ICD classification, and model interpretation into a unified framework. This review evaluates recent methodologies (2019–2024), highlights their strengths and limitations, and explores potential future developments. Ultimately, this review seeks to provide a comprehensive perspective on how NLP can address these challenges, improve EMR usability, and enhance the quality of healthcare delivery.

## 2. RESEARCH METHOD

This section outlines the procedures followed to conduct this review. The process begins by formulating research questions that address multiple key topics. Based on these questions, relevant keywords and literature are identified and searched. Subsequently, information from each source is extracted and analyzed to highlight the differences between the various approaches. Figure 1 illustrates this process, which consists of five essential steps. Each of these steps is explained in detail in the following subsections.



Figure 1. Review procedure

### 2.1. Research questions

Given the complexity of interpreting EMRs for abbreviation expansion, automatic diagnosis, and ICD classification, targeted research questions are essential. Clinical notes are often ambiguous and context-dependent, yet studies address these tasks separately. The first research question is:

- 1) Why does EMR analysis usually focus on tasks like abbreviation expansion, automatic diagnosis, or ICD classification separately, and not address all these areas together?

The second research question explores the key methods and requirements for efficient EMR deciphering:

- 2) What methods are used to decipher EMRs, such as expanding abbreviations, automatic diagnosis, and ICD classification, and what characteristics are needed for these methods to be successful?

Finally, identifying and overcoming challenges in NLP implementation is critical. The third research question is:

- 3) What are the main challenges in using NLP for EMR tasks like abbreviation expansion, automatic diagnosis, and ICD classification, and how can they be overcome?

These questions guide the review to enhance accuracy and efficiency in EMR deciphering.

## 2.2. Search keywords and literature search

This study used databases like IEEE, ScienceDirect, Google Scholar, Web of Science (WoS), and Scopus, with most papers sourced from Google Scholar. The selection criteria were:

- The method was developed or applied between 2019 and early 2024.
- Studies focusing on NLP applications in medical settings.
- Research on deciphering EMRs, clinical notes, or medical notes.
- Methods specific to abbreviation expansion, automatic diagnosis, or ICD classification.

Figure 2 illustrates the evolution of search keywords, including "auto diagnosis," "ICD classification," "deciphering clinical notes or medical notes," "NLP," and "deciphering EMRs," showing the search terms, results, and selected papers. We begin by searching for papers using these keywords and then select relevant literature, such as [23] on ICD classifications and [24] on expanding abbreviations in medical notes. Some articles were excluded for not meeting the criteria, including duplicates.

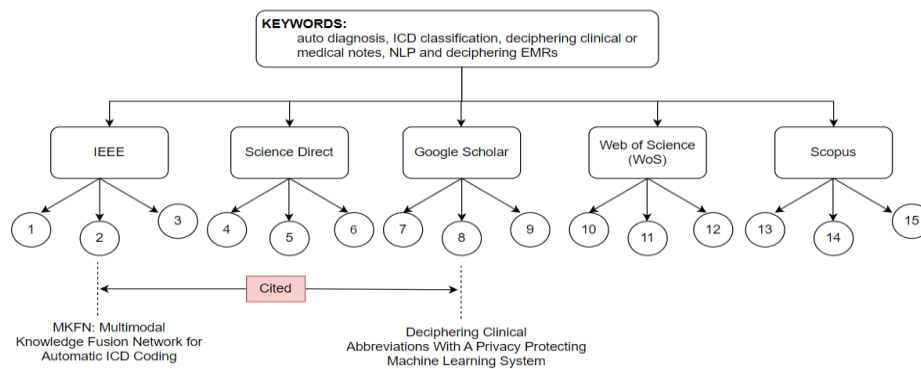


Figure 2. Search keywords and literature search procedure

## 2.3. Knowledge extraction and exclusion

The knowledge extraction process involved identifying relevant data from the selected studies, focusing on methods, motivations, and standards for EMR deciphering tasks like abbreviation expansion, automatic diagnosis, and ICD classification. Studies published between 2019 and 2024 were prioritized. Exclusion criteria were put in place to keep the review focused by removing irrelevant, outdated, or methodologically weak studies. The studies were critically analyzed to identify conflicting information and gaps in the literature. This study synthesizes the recent advances in NLP for medical informatics by integrating these processes, underlining key findings, and identifying gaps in current research. This will provide insights for future studies to develop integrated methodologies for deciphering EMRs.

## 3. CHALLENGES IN INTEGRATING EMR DECIPHERING, AUTOMATIC DIAGNOSIS AND ICD CODING

EMR deciphering significantly influences why most research tends to focus on deciphering EMRs like strategies for expanding abbreviations in EMR, automatic diagnosis, or ICD classification, rather than addressing them concurrently. A primary reason for this divide is the complexity of clinical language. EMRs are composed of unstructured clinical notes that incorporate medical jargon, abbreviations, and context-specific terms [5]. NLP is particularly effective for navigating the intricacies and nuances of human language, making it a vital tool for interpreting EMRs. To effectively preprocess unstructured text data in EMRs, various NLP techniques, such as tokenization, part-of-speech tagging, and named entity recognition, are employed [5], [25], [26].

These preprocessing steps are crucial for transforming raw clinical text into structured data that can be analyzed further [27]. Advanced NLP models, like BERT and ClinicalBERT [9], [8], utilize contextual

embeddings to grasp the meanings of words and phrases within the broader context of clinical notes [28], [29]. This contextual comprehension is essential for accurately interpreting medical information [30]. Furthermore, NLP models can be fine-tuned for specific tasks, such as expanding abbreviations, automatic diagnosis, or ICD coding [22], [31]. However, the unique requirements and challenges associated with each task necessitate specialized methodologies. This specialization often leads researchers to concentrate on a single task at a time to achieve optimal results. By focusing on one task, researchers can enhance their models and methods for that specific purpose, whether it involves deciphering EMRs or automating ICD coding [32].

The distinct challenges tied to each task clarify why research frequently emphasizes either EMR deciphering or ICD classification. EMR deciphering requires understanding and structuring unstructured or structured text, while ICD classification involves mapping clinical information to standardized codes [27], [33]. Each task presents unique hurdles and may require different methodologies and evaluation metrics. Moreover, models tailored for EMR deciphering are typically optimized for natural language understanding, whereas those designed for ICD classification focus on classification tasks. Attempting to combine these tasks can complicate the model training and evaluation processes.

In contrast, automatic diagnosis integrates different data, such as patient history, symptoms, and laboratory results. Most state-of-the-art NLP models are designed for purely text-based analysis and thus face challenges in integrating these types of data. Models need large, varied datasets to understand the gamut of possibilities in a medical context to make appropriate diagnoses. Insufficient data means a lack of specialized knowledge to grasp such complex conditions and arrive at dependable diagnostic decisions.

Finally, concentrating on one task allows researchers to allocate resources and time more effectively. Conducting comprehensive studies on both EMR deciphering and ICD classification simultaneously can be resource-intensive and may dilute efforts to achieve high performance in either task. By leveraging NLP for EMR deciphering, researchers can tackle the complexities of clinical language and enhance the accuracy of interpreting medical texts. This focus on specialization helps explain why most studies prioritize either EMR deciphering or ICD classifications, but not at the same time. Understanding these dynamics can inform future research directions toward more integrated and holistic approaches in the field of medical informatics.

## 4. NLP METHODS

### 4.1. Large language models

One notable method within NLP is the use of large language models (LLMs), which gained widespread attention with the emergence of ChatGPT. LLMs, including models like bidirectional encoder representations from transformers (BERT), generative pre-trained transformer (GPT), and text-to-text transfer transformer (T5), possess architectures that allow them to analyze context from both directions in a sequence of words. This capability significantly enhances their understanding of word meanings in context. These models are typically pre-trained on a wide array of tasks, such as language modeling, text completion, and other NLP-related functions. This pre-training phase equips the models with a solid foundation in general language patterns, making LLMs powerful tools for understanding and generating human-like text. They are invaluable for various language-related tasks and applications in healthcare [23], [34]-[37].

Moreover, transformer-based architectures like BERT and GPT have demonstrated exceptional performance. BERT's bidirectional context understanding enables it to grasp nuanced meanings in clinical text effectively, making it well-suited for tasks requiring detailed comprehension. GPT, renowned for its generative capabilities, can also be fine-tuned for specific applications such as automated disease diagnosis, albeit at the cost of increased computational resources [24], [36], [37]. Numerous previous studies have leveraged LLMs [27], [31] to expand abbreviations as a strategy for deciphering EMRs.

T5, developed by Google Research, introduces a unified framework for a range of NLP tasks by framing them as text generation problems. In this model, both input and output are treated as text strings, allowing for a consistent training approach across diverse tasks. By employing this text-to-text format, T5 leverages its pre-trained capabilities to generate contextually relevant and semantically meaningful medical text. When tasked with expanding abbreviations, T5 interprets the input as an abbreviated form and generates the corresponding expanded form or a detailed explanation, thereby facilitating clearer communication in clinical settings [6].

Reviews by [3] outlined that state-of-the-art clinical NERs are all variants of BERT, mostly fine-tuned or trained on domain-specific corpora. The very significant drawback with these models is the lack of variety in the medical corpus the models have been trained with, which might negatively impact performance. BERT models are extremely dependent on contextual information, and if the corpus is not representative of all medical specialties, contexts, and document types, generalization to diverse health scenarios can be poor. Even domain-

specific variants such as ClinicalBERT and BioBERT [8], [25], [30], while improving performance, have most of the same challenges associated with the representativeness of the corpus.

On the other hand, the pretraining of GPT models involves diverse general texts such as books, websites, and essays. This broad training data enables GPT to work on different kinds of tasks, but it lacks specificity in a particular domain. More specifically, for healthcare purposes, GPT models have to be finetuned. Regardless, GPT models, in health especially, promise to make quite a revolution in clinical decision support, improvement of communication with patients, and generally in smoothing data management processes [38]. However, domain-specific knowledge has proved to be a big challenge for GPT models, such as ChatGPT and GPT-4, whereas task-specific models outperform generalized models like ChatGPT [39].

In the evolving landscape of deep learning-based natural language processing, BERT models are recognized as state-of-the-art for various language understanding benchmarks. By utilizing a self-attention mechanism and transfer learning, BERT has outperformed previous models in numerous NLP downstream tasks. Pretrained BERT models are constructed on diverse text corpora; however, ClinicalBERT, a tailored variant of BERT, is specifically trained on specialized corpora that include clinical texts. This targeted training allows the representations learned from medical texts to be effectively processed for subsequent clinical applications [40].

LLMs like BERT, GPT, and T5 hold great potential for expanding abbreviations and improving EMR interpretation. To maximize their impact in healthcare, addressing limitations such as contextual understanding and domain knowledge is crucial. Enhancing performance through diverse training corpora, hybrid models, and fine-tuning for specific tasks can improve ICD classification, diagnosis, and EMR deciphering. However, without robust understanding, these models risk incorrect diagnoses and ICD coding, potentially compromising patient care.

#### 4.2. Information extraction methods

Information extraction (IE) is a core NLP technique that structures information from unstructured or semi-structured text. It is widely applied in healthcare to process EMRs, identifying key entities like diseases, medications, and procedures to enable efficient ICD classification and support tasks like automatic diagnosis. IE techniques such as named entity recognition (NER), LLMs, and hybrid models are instrumental in converting complex clinical notes into structured data [41]-[45].

A notable study applied IE for ICD-10 classification using co-occurrence analysis and embedding-based representations to extract synonyms, hyponyms, and hypernyms [4]. While effective for shorter n-grams, this approach struggled with longer sequences, requiring alignment with embedding spaces and disabling syntactic filters. This limitation reduces generalizability across contexts. Hybrid methods that integrate syntactic and semantic information may overcome such challenges and improve the extraction of longer terms [16].

Data mining techniques were applied to improve patient safety in classification, decision trees, and eigenvalue analysis. Techniques were scalable and performed on knowledge visualization, diagnosis improvement, and medication decisions [18], [42], [43]. Data mining models themselves often have issues such as interpretability, data sets that change distribution, and documents existing as unstructured data in handwritten notes [44]. Solutions may include transfer learning, coupled with more interpretable models to accommodate changing data patterns, hence making improvements in unstructured data processing.

A critical limitation of the current IE and data mining methods is their inability to handle the contextual ambiguity in clinical abbreviations. For example, "CHF" may indicate "congestive heart failure" or "chronic heart failure," depending on the context. Without sophisticated contextual analysis, models may misunderstand these abbreviations, leading to erroneous diagnoses and treatment decisions. Integrating contextual understanding into development techniques is crucial for accuracy improvement in abbreviation expansion and ultimately, in patient care. While IE offers several advantages in structuring clinical data and supporting diagnosis, limitations like handling of context, interpretability, and adaptability to unstructured data raise the challenges that need to be covered in further research to tune these methods for more robustness in healthcare applications.

Optical character recognition (OCR) has emerged as a robust technique for extracting data from EMRs and clinical notes, efficiently processing various document types, including handwritten and printed materials [15], [45]. Its rapid processing capabilities and high accuracy are attributed to advancements in machine learning and computer vision [46], [47]. However, OCR struggles with low-quality images and poor handwriting, leading to inaccuracies. To mitigate these challenges, improvements in preprocessing techniques and diverse training datasets can enhance OCR's robustness in real-world applications.

A hybrid method, hybrid-NER (hNER), integrates dictionary-based approaches and human-in-the-loop (HITL) validation to identify medical entities like symptoms and dosage forms [47]. The dictionary-based method ensures accuracy by comparing entities to predefined dictionaries, while HITL allows domain experts to validate predictions. Although HITL provides flexibility and insights for ambiguous entities, it is

limited by subjectivity and variability in expert interpretations. This limitation highlights the need for training and standardized guidelines to ensure consistency.

ClinicalBERT, a specialized BERT model, has been adapted for medical text processing, excelling in tasks like NER and medical concept identification [31]. Its multitask adaptation, Multitask-ClinicalBERT, addresses multiple clinical tasks simultaneously. However, its multitasking nature can dilute performance in specialized tasks like abbreviation expansion, which requires deep contextual understanding. Limitations also include its inability to adapt quickly to emerging medical terminology and practices. Incorporating explanation methods, such as interpretability techniques, may partially address these issues.

The development of automated ICD coding has evolved over the decades, with numerous researchers creating various methods to reduce the time-consuming tasks typically handled by human coders. One frequently employed approach is the rule-based methodology, which seeks to transform plain text into executable logical decisions for automatic coding prediction. A key aspect of enhancing the effectiveness of these rule-based systems is the inclusion of a broader range of medical concepts within the coding rules. By integrating equivalent terms from guidelines along with their synonyms, abbreviations, and related information, researchers can expand the vocabulary of health-related terms covered. However, rule-based approaches come with notable drawbacks, particularly in terms of flexibility and adaptability. The overlap of symptoms across multiple diseases can lead to issues such as over-coding and missed codes. Furthermore, as the number of codes increases, disputes may arise between conflicting rules, complicating the coding process [10], [17], [34], [37], [47], [48].

Mayya *et al.* [49] introduces the label attentive transformer architectures (LATA) model, which enhances input context learning for specific output classes in NLP tasks with limited training samples but numerous output classes. LATA automates ICD-10 code assignments using patient case reports and employs label attention mechanisms across BERT variants to improve predictive accuracy. A significant contribution of the study is the use of a unified tokenizer and consistent hyperparameters across BERT variants, enabling insights into parameter variations. LATA also addresses the need for explainable clinical decision support systems (CDSSs) by visualizing attention weights, thereby linking clinical note text to diagnostic codes and enhancing trust in the model. While the study does not explicitly address limitations, it highlights reducing false positives as a key future direction to improve coding accuracy and reliability in healthcare systems.

Another study aims to develop a predictive model for ICD codes using the MIMIC-III clinical text dataset. By leveraging natural language processing techniques and deep learning architectures, the researchers constructed a pipeline to extract relevant information from MIMIC-III, a large, de-identified, and publicly accessible medical records database. Their methodology predicts diagnosis codes from unstructured data, including discharge summaries and notes detailing symptoms. They employed state-of-the-art deep learning algorithms such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, bidirectional LSTM (BiLSTM), and BERT model, by first tokenizing the clinical text with Bio-ClinicalBERT, a pre-trained model from Hugging Face. To evaluate the effectiveness of their approach, the researchers conducted experiments using the discharge dataset from MIMIC-III. They explored a variety of deep learning models, with particular emphasis on Bio-ClinicalBERT, which is specifically pre-trained for biomedical texts [2].

They reported that by utilizing the BERT model, their approach achieved high accuracy in predicting the top 10 and top 50 diagnosis codes, which refer to the most frequently assigned diagnostic codes in the MIMIC-III dataset. This dataset contains de-identified health data, and predicting these top codes is crucial for automating medical coding tasks. This focus enhances the model's ability to accurately interpret clinical language. However, the study identifies several limitations, including the demands on computational resources and challenges related to the complexity and heterogeneity of healthcare data within MIMIC-III, which can result in issues such as imbalanced classes and missing values that negatively impact model performance. Lastly, they highlight the importance of addressing the limitations in explainability and transparency associated with these complex models [50].

#### 4.3. Explainable artificial intelligence

While explainable artificial intelligence (XAI) itself is a broader field focused on making AI models interpretable and understandable, it becomes especially relevant and integrated within the context of NLP when these models are used to process and analyze text. Thus, it can certainly consider XAI as part of the NLP landscape, particularly in applications where understanding model decisions is critical. Clear explanations are essential, especially in the medical field as confusion or ambiguity can have serious consequences. Misunderstandings in critical medical situations can lead to errors in patient care, diagnosis, or treatment [7], [15], [51]-[53]. For example, linking medical abbreviations to their expanded forms can reduce the chances of mistakes and improve the quality of care. In healthcare, where accuracy is vital, providing clear and understandable information is key to maintaining high standards.

Several XAI methods help interpret machine learning model outputs. Local Interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP) are two popular tools. LIME creates simpler models by locally approximating the behavior of complex models around specific instances, making it easier to understand the influence of features on predictions. SHAP, grounded in game theory, assigns credit to each feature by averaging contributions from all possible feature subsets. Both give clear explanations but are computationally expensive since they run the model multiple times; hence, they cannot be used for real-time explanations, especially for large neural networks [33]. LIME also approximates a locally linear model rather than directly explaining black-box models [54].

Another approach, integrated gradients (IG), is calculating word importance in a phrase by utilizing gradient-based techniques. It requires the model to be non-mandatory differentiable, so it works for complex models, including transformers. IG does this by calculating the gradient for both baseline and target inputs to assess the contribution of each word to the prediction. Although IG is very enlightening, one study demonstrated that LLMs not pre-trained on medical data produced poor performance on complex clinical indications, which led to incorrect predictions. Moreover, 15% of the errors were due to missing key input sections, probably induced by training biases or the lack of good edge cases.

Li *et al.* [46] performed a study that combined a LLM with semantic role labeling (SRL) to detect the main predicate, arguments, and their relations in a sentence. SRL works by labeling the grammatical roles of words, identifying predicates, and linking them to relevant arguments, such as patients or doctors. Although the proposed method performed well on the small dataset, performance increases with larger datasets, where more diverse examples are included. However, if applied to out-of-domain texts, SRL faces challenges that reduce its robustness and effectiveness. In all, these XAI methods enhance model interpretability. However, they all have their own shortfalls: computational complexity, susceptibility to training data, and limited performance on out-of-domain texts. While some of these can be mitigated through more diverse datasets or domain-specific training, others may indeed require further innovation to fully optimize their applicability in real-time and in a clinical setting.

Explainability is vital in the medical domain for clear communication among healthcare professionals and patients, particularly for clinical abbreviations [55]. Methods like LIME, SHAP, IG, and SRL enhance interpretability in NLP but face challenges. LIME and SHAP are computationally intensive, hindering real-time use, which approximation methods could address. IG struggles with complex medical language, requiring domain-specific training, while SRL's performance on out-of-domain texts could improve through adaptive learning. Biases in training data also pose risks, emphasizing the need for thorough auditing. Tackling these issues is crucial for reliable NLP models, ensuring precision in medical communication and patient care.

## 5. RESULT AND DISCUSSION

Implementing NLP techniques for EMR deciphering and ICD coding presents several key challenges. One significant issue is the lack of diverse training data; many NLP models, including BERT and its variants, rely heavily on domain-specific training datasets that often lack the necessary diversity. This limitation can hinder the model's ability to generalize across various medical specialties and types of clinical texts. To address this challenge, it is crucial to develop more comprehensive training datasets that encompass a wider range of medical contexts. Additionally, general-purpose models like GPT may struggle with tasks requiring specialized knowledge, as their pre-training data is not tailored to healthcare contexts. Fine-tuning these models on specific healthcare datasets can enhance their performance, while hybrid models that combine the strengths of different architectures such as BERT for understanding context and GPT for generating coherent text could improve their effectiveness in clinical applications.

Another challenge lies in the complexity of clinical language, where the nuanced and varied nature of medical terminology can hinder NLP systems' understanding, making it difficult to accurately interpret abbreviations and concepts. Implementing advanced techniques like domain adaptation and utilizing knowledge graphs to inform models about the relationships between medical terms can improve comprehension and accuracy. Furthermore, integrating NLP systems into existing healthcare workflows poses difficulties due to potential resistance to change and the need for user-friendly interfaces. Engaging stakeholders early in the development process and ensuring that NLP tools align with clinical needs will facilitate smoother adoption.

Lastly, regulatory and ethical concerns surrounding patient privacy and compliance with healthcare regulations are critical to consider. Developing clear guidelines for data usage, incorporating robust security measures, and ensuring transparency in model decisions can help address these concerns. By tackling these challenges through targeted data collection, fine-tuning, hybrid modeling, and careful integration into clinical workflows, we can significantly enhance the effectiveness of NLP techniques for EMR deciphering and ICD coding, ultimately leading to improved patient outcomes and streamlined healthcare processes.



## 6. CONCLUSION

In conclusion, the implementation of NLP techniques for EMR deciphering and ICD coding holds significant promise for enhancing healthcare delivery. However, the key challenges associated with diverse training data, the complexity of clinical language, integration into existing workflows, and regulatory compliance must be effectively addressed to maximize the potential of these technologies. By developing comprehensive training datasets that encompass a broad range of medical contexts and employing fine-tuning strategies for specialized models, we can improve the accuracy and reliability of NLP applications in clinical settings. Furthermore, leveraging hybrid models that combine the strengths of various architectures can lead to better understanding and generation of medical texts, facilitating clearer communication and improved decision-making. Engaging stakeholders throughout the process and ensuring compliance with ethical and regulatory standards will promote the successful adoption of NLP tools in healthcare. Ultimately, by overcoming these challenges, we can harness the full capabilities of NLP to streamline clinical workflows, improve patient outcomes, and transform the landscape of healthcare analytics and documentation.

## ACKNOWLEDGEMENTS

We extend our heartfelt gratitude to our research team for their unwavering support, guidance, and invaluable feedback throughout this project. We also thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for providing the essential resources for this study. We appreciate the contributions of the researchers and authors reviewed in this paper. This work was supported by the Higher Institution Centre of Excellence (HICoE) research grant 600-RMC/MOHE HICoE CARE-I 5/3 (01/2025) awarded to the Cardiovascular Advancement and Research Excellence Institute (CARE Institute), Universiti Teknologi MARA. Thank you all for making this achievement possible.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nurul Anis Balqis Iqbal Basheer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Sharifalillah Nordin	✓							✓	✓	✓		✓	✓	
Sazzli Shahlan Kassim	✓			✓	✓		✓	✓		✓		✓	✓	✓
Azliza Mohd Ali							✓		✓	✓				
Nurzeatul Hamimah									✓	✓				
Abdul Hamid														

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

## REFERENCES

- [1] A. Bohr and K. Memarzadeh, "The rise of artificial intelligence in healthcare applications," in *Artificial Intelligence in Healthcare*, 2020, pp. 25–60. doi: 10.1016/b978-0-12-818438-7.00002-2.
- [2] H. Goodrum, K. Roberts and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," *International Journal of Medical Informatics*, vol. 144, p. 104302, Dec. 2020, doi: 10.1016/j.ijmedinf.2020.104302.






- [3] M. C. Durango, E. A. Torres-Silva and A. Orozco-Duque, "Named entity recognition in electronic health records: a methodological review," *Healthcare Informatics Research*, vol. 29, no. 4, pp. 286–300, Oct. 2023, doi: 10.4258/hir.2023.29.4.286.
- [4] J. Ferguson, "ICD-10 Procedure codes: harnessing the power of procedure codes," *Health Catalyst*, 2023. [Online]. Available: <https://www.healthcatalyst.com/insights/icd-10-pcs-harnessing-the-power-of-procedure-codes> (accessed 8 October 2024).
- [5] F. Teng, Y. Liu, T. Li, Y. Zhang, S. Li and Y. Zhao, "A review on deep neural networks for ICD coding," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 5, pp. 4357–4375, 1 May 2023, doi: 10.1109/TKDE.2022.3148267.
- [6] C. Blease, J. Torous and M. Hägglund, "Does patient access to clinical notes change documentation?," *Frontiers in Public Health*, vol. 8, Nov. 2020, doi: 10.3389/fpubh.2020.577896.
- [7] H.-J. Kong, "Managing unstructured big data in healthcare system," *Healthcare Informatics Research*, vol. 25, no. 1, pp. 1–2, Jan. 2019, doi: 10.4258/hir.2019.25.1.1.
- [8] V. D. Lai *et al.*, "ChatGPT Beyond English: towards a comprehensive evaluation of large language models in multilingual learning," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2304.05613.
- [9] A. M. Nair and K. R. Bindu, "Semantic role labelling using transfer learning model," *Journal of Physics Conference Series*, vol. 1767, no. 1, p. 012024, Feb. 2021, doi: 10.1088/1742-6596/1767/1/012024.
- [10] C. Basu, R. Vasu, M. Yasunaga, and Q. Yang, "Med-EASI: finely annotated dataset and models for controllable simplification of medical texts," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2302.09155.
- [11] C. Lu, C. K. Reddy, P. Wang, and Y. Ning, "Towards semi-structured automatic ICD coding via tree-based contrastive learning," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2310.09672.
- [12] C. Pesquita, "Towards semantic integration for explainable artificial intelligence in the biomedical domain," *Science and Technology Publications*, Jan. 2021, doi: 10.5220/0010389707470753.
- [13] N. S. Faramarzi, M. Patel, S. H. Bandrupally, and R. Banerjee, "Context-aware medication event extraction from unstructured text," *Aclanthology*, Jan. 2023, doi: 10.18653/v1/2023.clinicalnlp-1.11.
- [14] A. Kugic, B. Pfeifer, S. Schulz, and M. Kreuthaler, "Embedding-based terminology expansion via secondary use of large clinical real-world datasets," *Journal of Biomedical Informatics*, vol. 147, p. 104497, Nov. 2023, doi: 10.1016/j.jbi.2023.104497.
- [15] A. Sathyan, A. I. Weinberg and K. Cohen, "Interpretable AI for bio-medical applications," *Complex Engineering Systems*, vol. 2, no. 4, p. 18, Jan. 2022, doi: 10.20517/ces.2022.41.
- [16] A. Millan-Fernandez-Montes, D. Perez-Rey, G. Hernandez-Ibarburu, M. B. Palchuk, C. Mueller, and B. Claerhout, "Mapping clinical procedures to the ICD-10-PCS: the German operation and procedure classification system use case," *Journal of Biomedical Informatics*, vol. 109, p. 103519, Sep. 2020, doi: 10.1016/j.jbi.2020.103519.
- [17] R. L. Johnson, H. Hedegaard, E. S. Pasalic, and P. D. Martinez, "Use of ICD-10-CM coded hospitalisation and emergency department data for injury surveillance," *Injury Prevention*, vol. 27, no. Suppl 1, pp. i1–i2, Mar. 2021, doi: 10.1136/injury-prev-2019-043515.
- [18] G. B. Gebremeskel, B. Hailu and B. Biazen, "Architecture and optimization of data mining modeling for visualization of knowledge extraction: Patient safety care," *Journal of King Saud University. Computer and Information Sciences*, vol. 34, no. 2, pp. 468–479, Feb. 2022, doi: 10.1016/j.jksuci.2019.12.001.
- [19] H. A. Shamsi, A. G. Almutairi, S. A. Mashrafi and T. A. Kalbani, "Implications of language barriers for healthcare: a systematic review," *Oman Medical Journal*, vol. 35, no. 2, p. e122, Mar. 2020, doi: 10.5001/omj.2020.40.
- [20] L. Ou, Y. Yao, X. Luo, X. Li and K. Chen, "ContextAD: context-aware acronym disambiguation with siamese BERT network," *International Journal of Intelligent Systems*, vol. 2023, pp. 1–14, Jul. 2023, doi: 10.1155/2023/5014355.
- [21] S. L. Fleming *et al.*, "MedAlign: a clinician-generated dataset for instruction following with electronic medical records," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2308.14089.
- [22] S. Talebi, E. Tong and M. R. K. Mofrad, "Exploring the performance and explainability of BERT for medical image protocol assignment," *medRxiv (Cold Spring Harbor Laboratory)*, Apr. 2023, doi: 10.1101/2023.04.20.23288684.
- [23] S. Wang, H. Lin, Y. Zhang, X. Li, and W. Qu, "MKFN: multimodal knowledge fusion network for automatic ICD coding," *IEEE International Conference on Bioinformatics and Biomedicine*, Dec. 2023, doi: 10.1109/bibm58861.2023.10385669.
- [24] A. Rajkomar *et al.*, "Deciphering clinical abbreviations with a privacy-protecting machine learning system," *Nature Communications*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-35007-9.
- [25] A. Rahman *et al.*, "Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges, and open issues," *Cluster Computing*, vol. 26, no. 4, pp. 2271–2311, Aug. 2022, doi: 10.1007/s10586-022-03658-4.
- [26] J. S. Alpert, "The electronic medical record in 2016: advantages and disadvantages," *Digital Medicine*, vol. 2, no. 2, pp. 48–51, Apr. 2016, doi: 10.4103/2226-8561.189504.
- [27] A. Soleimani, C. Monz and M. Worring, "BERT for evidence retrieval and claim verification," *Advances in Information Retrieval*, 2020, pp. 359–366. doi: 10.1007/978-3-030-45442-5\_45.
- [28] Y. Kim, J.-H. Kim, Y.-M. Kim, S. Song and H. J. Joo, "Predicting medical specialty from text based on a domain-specific pre-trained BERT," *International Journal of Medical Informatics*, vol. 170, p. 104956, Feb. 2023, doi: 10.1016/j.ijmedinf.2022.104956.
- [29] J. Ye *et al.*, "The roles of electronic health records for clinical trials in low- and middle-income countries: scoping review," *JMIR Medical Informatics*, vol. 11, 2023, doi: 10.2196/47052.
- [30] S. B. Atallah, N. R. Banda, A. Banda and N. A. Roek, "How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery," *Techniques in Coloproctology*, vol. 27, no. 8, pp. 609–614, Jul. 2023, doi: 10.1007/s10151-023-02837-8.
- [31] A. Mulyar, O. Uzuner and B. McInnes, "MT-clinical BERT: scaling clinical information extraction with multitask learning," *Journal of the American Medical Informatics Association*, vol. 28, no. 10, pp. 2108–2115, Aug. 2021, doi: 10.1093/jamia/ocab126.
- [32] P. N. Ngugi, M. C. Were and A. Babic, "Users' perception on factors contributing to electronic medical records systems use: a focus group discussion study in healthcare facilities setting in Kenya," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01737-x.
- [33] V. Alonso *et al.*, "Health records as the basis of clinical coding: Is the quality adequate? A qualitative study of medical coders' perceptions," *Health Information Management Journal/Health Information Management*, vol. 49, no. 1, pp. 28–37, Feb. 2019, doi: 10.1177/1833358319826351.
- [34] J. Park, "Selectively connected self-attentions for semantic role labeling," *Applied Sciences*, vol. 9, no. 8, p. 1716, Apr. 2019, doi: 10.3390/app9081716.
- [35] R. Abeyasinghe *et al.*, "Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method," *Journal of Biomedical Informatics*, vol. 134, p. 104162, Oct. 2022, doi: 10.1016/j.jbi.2022.104162.
- [36] S. Sung, H.-A. Park, H. Jung, and H. Kang, "A SNOMED CT mapping guideline for the local terms used to document clinical findings and procedures in electronic medical records in South Korea: methodological study," *JMIR Medical Informatics*, vol. 11, p. e46127, Apr. 2023, doi: 10.2196/46127.




- [37] T. Santos *et al.*, "PathologyBERT -- Pre-trained vs. a new transformer language model for pathology domain," *arXiv (Cornell University)*, doi: 10.48550/arxiv.2205.06885.
- [38] M. Wornow *et al.*, "The shaky foundations of large language models and foundation models for electronic health records," *Npj Digital Medicine*, vol. 6, no. 1, Jul. 2023, doi: 10.1038/s41746-023-00879-8.
- [39] Y. Liu *et al.*, "Summary of ChatGPT-related research and perspective towards the future of large language models," *Meta-radiology*, vol. 1, no. 2, p. 100017, Sep. 2023, doi: 10.1016/j.metrad.2023.100017.
- [40] N. Saraswat, C. Li and M. Jiang, "Identifying the question similarity of regulatory documents in the pharmaceutical industry by using the recognizing question entailment system: evaluation study," *JMIR AI*, vol. 2, p. e43483, Sep. 2023, doi: 10.2196/43483.
- [41] Z. Shuai *et al.*, "Comparison of different feature extraction methods for applicable automated ICD coding," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, Jan. 2022, doi: 10.1186/s12911-022-01753-5.
- [42] L. J. Y. Flores, H. Huang, K. Shi, S. Chheang, and A. Cohan, "Medical text simplification: optimizing for readability with unlikelyhood training and reranked beam search decoding," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2310.11191.
- [43] N. Joseph *et al.*, "Automated data extraction of electronic medical records: validity of data mining to construct research databases for eligibility in gastroenterological clinical trials," *Upsala Journal of Medical Sciences*, vol. 127, Jan. 2022, doi: 10.48101/ujms.v127.8260.
- [44] D. H. P. Benicio, J. C. Xavier-Júnior, K. R. S. De Paiva and J. D. De Araújo Sant Camargo, "Applying text mining and natural language processing to electronic medical records for extracting and transforming texts into structured data," *Social Science Research Network*, Jan. 2021, doi: 10.2139/ssrn.3991515.
- [45] S. N. Laique *et al.*, "Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports," *Gastrointestinal Endoscopy*, vol. 93, no. 3, pp. 750–757, Mar. 2021, doi: 10.1016/j.gie.2020.08.038.
- [46] X. Li, H. Chen, C. Liu, J. Li, M. Zhang, J. Yu and M. Zhang, "LLMs can also do well! Breaking barriers in semantic role labeling via large language models," *arXiv*, 2025, doi: 10.48550/arXiv.2506.05385.
- [47] R. Ramachandran and K. Arutchelvan, "Named entity recognition on bio-medical literature documents using hybrid based approach," *Journal of Ambient Intelligence & Humanized Computing*, Mar. 2021, doi: 10.1007/s12652-021-03078-z.
- [48] D. Weatherspoon and A. Chattopadhyay, "International classification of diseases codes and their use in dentistry," PubMed Central (PMC), 2013. [Online]. Available: <https://ncbi.nlm.nih.gov/pmc/articles/PMC4394630/> (accessed 8 October 2024)..
- [49] V. Mayya, S. S. Kamath and V. Sugumaran, "LATA - label attention transformer architectures for ICD-10 coding of unstructured clinical notes," *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 2021, doi: 10.1109/cibcb49929.2021.9562815.
- [50] I. Aden, C. H. T. Child and C. C. Reyes-Aldasoro, "International classification of diseases prediction from MIMIC-III clinical text using pre-trained clinicalBERT and NLP deep learning models achieving state of the art," *Big Data and Cognitive Computing*, vol. 8, no. 5, p. 47, 2024, doi: 10.3390/bdcc8050047.
- [51] H. Dong, V. Suárez-Paniagua, W. Whiteley and H. Wu, "Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation," *Journal of Biomedical Informatics*, vol. 116, p. 103728, Apr. 2021, doi: 10.1016/j.jbi.2021.103728.
- [52] J. Amann, A. Blasimme, E. Vayena, D. Frey and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Nov. 2020, doi: 10.1186/s12911-020-01332-6.
- [53] P. Zhang and M. N. K. Boulos, "Generative AI in medicine and healthcare: promises, opportunities and challenges," *Future Internet*, vol. 15, no. 9, p. 286, Aug. 2023, doi: 10.3390/fi15090286.
- [54] S. Mirzaei, H. Mao, R. R. O. Al-Nima, and W. L. Woo, "Explainable AI evaluation: a top-down approach for selecting optimal explanations for black-box models," *Information*, vol. 15, no. 1, p. 4, Dec. 2023, doi: 10.3390/info15010004.
- [55] S. Ali *et al.*, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

## BIOGRAPHIES OF AUTHORS






**Nurul Anis Balqis Iqbal Basheer**    is pursuing her master's degree in computer science at Universiti Teknologi MARA (UiTM), Shah Alam. She obtained her diploma in computer science from UiTM, Kedah. Specializing in artificial intelligence, she has a keen focus on intelligence programming and data science. She concentrated on predicting stock prices during her bachelor's studies at UiTM, Shah Alam. Currently, her research interest lies in bioinformatics. She can be contacted at email: [nurulanisbalqisiqbalbasheer@gmail.com](mailto:nurulanisbalqisiqbalbasheer@gmail.com).






**Dr. Sharifalillah Nordin**    learned her Bachelor of Information Technology from Universiti Utara Malaysia (UUM) in 2001, followed by a Master of Science in Internet Computing from the University of Surrey, UK, in 2003, and a Ph.D. in bioinformatics from Universiti Malaya (UM) in 2010. Currently serving as a Senior Lecturer at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, she has been a member of the UiTM academic community since 2009. She is deeply committed to teaching and research, with her expertise spanning biodiversity informatics, knowledge engineering, and artificial intelligence. For inquiries, she can be contacted at email: [sharifalillah@uitm.edu.my](mailto:sharifalillah@uitm.edu.my).






**Prof. Dr. Sazzli Shahlan Kassim**    is the National Specialist Register (NSR) Certified trainer in Cardiology. He also leads the construction of Universiti Teknologi MARA's teaching hospital in Puncak Alam, Selangor. With a passion for raising the standards of cardiology care and services in Malaysia, he has been actively involved in the education and training of the medical community in this field. He has delivered many lectures and seminars for both the public and healthcare professionals, raising awareness of cardiovascular conditions and other disease areas. He can be contacted at email: sazzlishahlan@uitm.edu.my.



**Dr. Azliza Mohd Ali**    received both her bachelor's degree in information technology (2001) and master's degree in intelligent knowledge-based systems (2003) from Universiti Utara Malaysia (UUM). She joined Universiti Teknologi MARA (UiTM) as a lecturer in 2004 and earned her Ph.D. in computer science from Lancaster University, UK. She is dedicated to university teaching and research, with current research interests in anomaly detection, data mining, machine learning, and knowledge-based systems. She can be contacted at email: azliza794@uitm.edu.my.



**Nurzeatul Hamimah Abdul Hamid**    is a senior lecturer of information systems at Universiti Teknologi MARA (UiTM). She obtained her master's degree in intelligent systems from the University of Sussex, UK, in 2005. She teaches courses on the fundamentals of artificial intelligence, AI programming paradigms, and intelligent agents. Her main research interests include software agents, normative multi-agent systems, and trust and reputation systems. She can be contacted at email: nurzea818@uitm.edu.my.